BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

MOLECULAR BREEDING: THE NATURAL APPROACH TO PROTEIN DESIGN

By JON E. NESS, STEPHEN B. DEL CARDAYRÉ, JEREMY MINSHULL, and WILLEM P. C. STEMMER

Maxygen, Redwood City, California 94063

The breeding of wild plants and animals into the agricultural domesticates we know today precedes Mendel's genetic studies by many millennia. Despite their lack of a formalized understanding, our ancestors harnessed the evolutionary power of sexual recombination of preexisting natural diversity to produce plants and animals with characteristics better suited to their needs. A recent advance in protein design, termed molecular breeding, allows protein engineers to homologously recombine multiple related genes by a process that closely mimics sexual recombination to generate functionally diverse libraries of chimeric proteins from which improved variants can be selected. Molecular breeding effects the permutation of diversity within a pool of related sequences and has proven to be an extraordinarily effective method to evolve proteins and pathways for better function.

I. Introduction

Just look around. Look at what nature has provided—things for basic necessities, such as the corn in breakfast cereal and cotton for clothing; things that bring pleasure, such as a pet dog or flowers. Most of these are not at all like what they were 5000 years ago. In fact, many have changed dramatically in only the last 100 years. This was accomplished by a process of selective breeding—a process as simple as mating one's two favorite dogs or saving the cream of the crop for next spring's

ADVANCES IN PROTEIN CHEMISTRY, Vol. 55

0065-3233/00 $35.00

planting. Of breeding, Darwin states: "The key is man's power of accumulative selection: nature gives successive variations; man adds them up in certain directions useful to him" (Darwin, 1859). Apart from the nature of selective pressure, Darwin recognized that the principles operating in the domestication and breeding of plants and animals were indistinguishable from those operating in natural selection. He recognized that natural variation within a population is the base material on which evolution and breeding depend. In addition to the similarity between breeding and natural selection, Darwin recognized that there must be a mechanism generating variation within the population from which nature or a breeder selects.

The block-wise exchange of homologous regions of chromosomes, which occurs during meiosis in sexually reproducing organisms (Roeder, 1997) and during genetic exchange in bacteria (Matic *et al.*, 1995; Ogunseitan, 1995), is by far the major generator of diversity. Preexisting diversity present within a population of organisms is shuffled to produce new variants. The fitter or more desirable organisms are those bearing a better mixture of beneficial alleles and fewer deleterious alleles; these are selectively propagated in nature or by a breeder. Accumulation of beneficial alleles over multiple generations under selective pressure, be it natural or imposed, can result in remarkable modifications from ancestral organisms. The role of homologous DNA recombination in accelerating evolution is evidenced by its chimeric signature in genes that have endured selective pressure. For example, the penicillin-binding proteins of β-lactam resistant *Neisseria* species show complex patterns of recombination between at least six ancestral species, differing in DNA sequence identity by up to 23% (Smith *et al.*, 1991). The diversity of life, both extinct and extant, and the rapidity with which microorganisms evolve (e.g., see van der Meer *et al.*, 1992) are testaments to the effective mechanisms nature has developed and breeders have tapped to improve useful plants and animals (Burbank *et al.*, 1914). Recognizing the power of selective breeding, protein engineers have captured sexual recombination in the test tube to rapidly improve the products of genes. The highly active, functionally diverse gene libraries generated by molecular breeding[1] have extended directed evolution to a plethora of proteins for which only limited throughput screens are feasible (Chang *et al.*, 1999; Christians *et al.*, 1999; Minshull and Stemmer 1999; Ness *et al.*,

[1] Maxygen is the leader in the emerging field of directed molecular evolution, the process by which novel genes are generated for commercial and research purposes. Maxygen's proprietary technologies, known as MolecularBreeding™, mimic the natural process of evolution and bring together advances in classical breeding, molecular biology, and genomics.

1999; Patten *et al.*, 1997; Zhang *et al.*, 1997). Molecular breeding has also been extended beyond proteins to pathways and viruses, as well as partial genomes (Crameri *et al.*, 1997; Soong *et al.*, 1999).

II. The Need for Better Proteins

Our ancestors learned to exploit a variety of natural enzymatic transformations for the production of food and drink, such as the malting of barley and its fermentation to beer. Today enzymes have found many applications. Examples include the multi-enzymatic production of high fructose corn syrup (Bhosale *et al.*, 1996; Hagedorn and Kaphammer, 1994), the synthesis of acrylamide (Kobayashi *et al.*, 1992), and the use of streptokinase (Wong *et al.*, 1994; Wu *et al.*, 1998) for the therapeutic degradation of arterial blood clots. There are many benefits from using enzymes. In contrast to chemical catalysts, enzymes are nontoxic, biodegradable, and can be produced by fermentations of cheap, renewable feedstocks. In chemical synthesis, the exquisite specificity (substrate-, regio-, and stereo-selectivity) and efficiency of enzymes can obviate the need for protecting and deprotecting groups and simplify end-product purification. In addition, enzymatic processes generally require more moderate reaction conditions and simpler, more flexible production units. The use of biocatalysts (both enzymes and whole cells) as benign alternatives to chemical catalysts for producing renewable chemicals, pharmaceuticals, polymers, and fuels is integral to realizing current visions of sustainable development (Nedwin, 1997).

Despite these advantages, there are discrete performance limitations that have impeded biocatalysts from realizing their industrial potential. These limitations originate in the natural physiological role that biocatalysts play. Most enzymes have evolved to function optimally on a narrow set of substrates and under the precise conditions (temperature, ionic strength, pH, etc.) of their natural niche. The efficiency of enzymes allows for their transient use and for a cell to recycle their amino-acid components, thus wild-type enzymes often are unstable and prone to chemical and biological degradation. Although these characteristics provide a benefit to a cell in nature, they are suboptimal for the biotechnologist hoping to exploit a biocatalyst. The substitution of an enzyme for a chemical process often requires that the protein is active for many hours, capable of functioning under reaction conditions alien to its natural milieu (such as extremes of temperatures and pH or in the presence of organic solvents), able to use nonnatural substrates, and cost effective relative to alternative chemical catalysts.

Proteins are also used clinically to treat a variety of diseases. Erythropoietin stimulates erythrocyte production in kidney dialysis and chemotherapy patients. Granulocyte stimulating factor enhances immune systems compromised by cancer treatments. Cytokines such as interferons and interleukins are used for their anti-viral and anti-tumor activities. Other proteins are used to inhibit or stimulate blood clotting. For the most part, the pharmaceutical protein industry relies on cloning native human genes and expressing and purifying their products in recombinant form.

As with proteins used as biocatalysts, natural human proteins may require optimization for use as pharmaceuticals. One frequent limitation is protein half-life. Another is selectivity. Cytokines often bind to multiple receptors on different cell types, with different physiological responses resulting from the different binding events. Thus, while a therapeutically relevant dose of a protein pharmaceutical may activate a targeted biological pathway, it may also increase harmful or even fatal side effects due to activation of secondary pathways. In this case, a cytokine that activated only one of the pathways would clearly be desirable. Another limitation of naturally occurring proteins is that they often require high doses (milligrams to grams), making their use as therapeutics economically impractical. Yield of active recombinant protein production in a heterologous system is another property that could be improved, as was done for the expression of the jellyfish green fluorescent protein (GFP) in *Escherichia coli* (Crameri *et al.*, 1996). The strategies described below are most frequently related to enzyme activity, yet they clearly also apply to modifying the properties of interest in proteins for pharmaceutical and a plethora of other uses.

III. Strategies for Optimizing Proteins

Natural selection and classical breeding both employ the same empirical strategy of creating variants and selecting those that perform best—that is the essence of all protein tailoring methods. They differ only in the sources of sequence variation and the methods by which this diversity is tested.

A. *Natural Diversity*

The most reasonable source of starting points is nature. If one protein does not perform exactly as required, perhaps a homolog isolated from a different organism will. For example, an enzyme that functions well in high salt conditions may best be isolated from a halophile. On the other hand, properties such as lack of product inhibition or ability to function in an organic solvent may never be found by searching through nature, simply because a cell's survival has never depended on these characteristics. Naturally occurring proteins can be thought of as analogous to the wild ancestors of modern crops or domesticated animals: good starting points, but unlikely to possess the full range of properties required for human purposes (Diamond, 1997). A different source of variants is to take an available protein whose function most closely approximates that desired, and make and test variations on that sequence.

B. *Variations on a Single Sequence*

1. *Rational Design*

Structure-based protein design relies on knowing the structure of a protein and on tools for molecular modeling to predict favorable amino acid changes. A small number of promising modifications are introduced into the gene by methods such as oligonucleotide-directed mutagenesis (Kunkel *et al.*, 1991), and the effects tested. This method tests only a small number of variations. However, our understanding of protein function, structure, folding, and interaction is still sufficiently imprecise that no matter how good the available structural data, the variants frequently fail to show the desired improvements. Although the approach has proven fruitful and general strategies for tailoring a few properties are slowly being illuminated, the approach is impeded by assumptions that discount the complexity of biological systems (Rubingh, 1997). Although recent advances for visualizing proteins as the dynamic structures they are (Arnold and Ornstein, 1997) and for the *de novo* design of protein folds (Dahiyat and Mayo, 1997) hint at a future when proteins may be built to desired specifications, the prospect of designing proteins for specific functions is still a long way off.

2. *Random Point Mutations*

Iterated random point mutagenesis coupled with a screen or selection to evolve an improved protein (Arnold, 1998a; Shao and Arnold, 1996) is a strategy for molecular evolution that stems from classical strain improvement of industrial microorganisms (described below). Random point mutagenesis of the gene for the protein to be improved is typically performed by error-prone PCR (Arnold, 1998a; Cadwell and Joyce, 1992; Cadwell and Joyce, 1994; Chen and Arnold, 1991; You and Arnold, 1996), but exposure to chemical mutagens (Taguchi *et al.*, 1998), or

mutator strains (Bornscheuer *et al.*, 1998; Low *et al.*, 1996) have also been used. Because most mutations are detrimental or neutral (Shafikhani *et al.*, 1997; Suzuki *et al.*, 1996), a low mutation frequency is employed to generate approximately one to two amino acid changes per protein. A sample of the library is subjected to an appropriate selection or screen to identify those variants that have the desired improvements. The process is repeated with the single best performer, often with increasingly stringent selection pressure or screening criteria. The result is the "asexual" evolution of an improved protein by stepwise accumulation of single mutations. The advantage of this method, in contrast to structure-based engineering or random cassette mutagenesis, is that little or no information regarding protein structure and function is required, and few assumptions are made. Further, the process generally provides new protein sequences with measurable improvements in a desired activity in relatively small libraries ($\sim 10^4$/iteration). Amino acid changes contributing to an evolved phenotype are frequently scattered throughout the protein sequence; changes whose effects could not have been accurately predicted or calculated (Arnold, 1998a; Spiller *et al.*, 1999). The unbiased nature of the process affects diverse aspects of protein activity that are not easily modeled, such as transcription, translation, protein folding, and protein-protein interaction with host systems. Individual clones commonly have improvements in several of these properties. Despite the success of iterated mutagenesis compared to rational approaches, the process is still impeded by the low quality of random mutations and the inherent limitations of an asexual evolutionary process that accumulates mutations at a rate of only about one per cycle (discussed below).

3. Oligonucleotide Randomization

A feature of random point mutagenesis by the methods described above is that it is limited to single point mutations. Only one nucleotide in any codon is usually changed, which means that on average less than six amino acids may be accessed (Jespers *et al.*, 1997) (the remaining fourteen amino acids could only be accessed by double- or triple-base mutations). Because the amino acids that are accessible with a single base mutation tend to be conservative replacements for the original residue, this important bias results in mutations that are more conservative than fully random. However, saturation mutagenesis is a way to circumvent this bias, for example, by using a cassette containing a randomized codon to replace the native codon with an equal mixture of all possible codons (Reidhaar-Olson *et al.*, 1991). Because this approach is the most mutagenic, it can only be applied to a small part of a protein

and is used primarily when enough structural information is available to target a certain area, but not enough for designing and testing specific single mutations. This approach therefore lies between the purely structural-based and random approaches. Molecular modeling and other rational approaches are employed to identify regions of a gene to target by random mutagenesis. Screening the mutant clones for variants of desired function identifies random mutational solutions lying within the targeted region. This approach has found success (Black *et al.*, 1996), but remains limited by several factors. These include its imposition of rational criteria, the ability to identify critical, contiguous regions for mutagenesis, the fact that beneficial mutations are often found throughout a polypeptide, and the low quality of random mutations that requires the testing of large libraries when multiple random changes are introduced into a critical region, such as the substrate binding pocket of an enzyme. The latter limitation can be reduced by using a biased randomization, for example, 70% of the wild-type base and 10% of each of the other bases (Reidhaar-Olson *et al.*, 1991) or doping for desired subsets of amino acids (Arkin and Youvan, 1992). Variations of this method include scanning saturation mutagenesis (Chen *et al.*, 1999) and recursive ensemble mutagenenesis (Delagrave *et al.*, 1993; Delagrave and Youvan, 1993). These processes all increase the genetic variation accessible at a codon (or contiguous series of codons); however, the methods require construction of many libraries for complete scanning of a protein and are difficult to iterate.

C. Recombination of Diversity

A fundamental difference between the mutagenesis methodologies previously described and the examples of natural evolution and breeding described in the Introduction (Section I) is the role played by recombination. Homologous recombination has two major effects on an evolving system. First, it avoids reinventing the wheel: Once a mutation that confers an improved phenotype has entered a population, recombination allows it to be tested in combination with all other beneficial mutations, rather than all of these combinations having to be derived *de novo*. Second, it allows efficient removal of deleterious mutations by replacing them with wild-type sequence, thereby avoiding the downward spiral of fitness resulting from deleterious mutations in asexual populations (Muller's ratchet) (Muller, 1964). Not only is recombination a feature of all biological systems, but computational simulations of recombination show dramatic increases in the speed and range of evolution when recombination is included in the algorithms (Forrest, 1993; Gibson,

1989; Holland, 1975; Kelly, 1994). Molecular breeding (also called DNA shuffling) was developed to mimic this essential feature of natural evolution. The sexual process of molecular breeding (Stemmer, 1994a; Stemmer, 1994b) has supplanted iterated random mutagenesis as the most efficient and rapid method for directing the evolution of nucleic acids and proteins.

The most widely used format for molecular breeding is *in vitro* fragmentation and reassembly of DNA (Fig. 1). In this format, DNA from a pool of selected mutants is randomly fragmented (e.g., with DNase I) and reassembled in a primerless DNA amplification reaction. The reassembly reaction is recombinogenic because fragments from one DNA sequence can prime homologous regions of different DNA sequences by template switching. In addition, the level of mutagenesis can be adjusted with the appropriate choice of DNA polymerase and reaction conditions (Zhao and Arnold, 1997c).

1. Recombination of Single Sequences by DNA Shuffling

The power of the combinatorial nature of DNA shuffling was first demonstrated using the TEM-1 β-lactamase (Stemmer, 1994b). When expressed in *E. coli*, the TEM 1 β-lactamase provides low-level resistance to the poorly hydrolyzed antibiotic cefotaxime. Three cycles of DNA shuffling (error-prone, to create the initial diversity), selection, and pooling of the most cefotaxime resistant mutants improved resistance from a minimal inhibitory concentration (MIC) of 0.02 μg/ml to 320 μg/ml for a 16,000-fold increase in resistance. Two rounds of backcrossing (i.e., shuffling with a molar excess of the parental gene) resulted in a mutant that was 32,000-fold more drug resistant. Backcrossing (Stemmer, 1994b) of genes that encode improved variants is a method to identify beneficial mutations while flushing out neutral and deleterious mutations (Zhao and Arnold, 1997b). In addition to improved turnover, this mutant retained a promoter mutation that resulted in a twofold increase in expression over the parental enzyme, illustrating the ability of DNA shuffling to solve a problem through more than one route.

This example demonstrates both of the recombination-derived advantages of DNA shuffling. The first advantage is that DNA shuffling can create combinations of distant, separately selected residues. In a previous approach to increase cefotaxime resistance by molecular modeling of the known structure of the TEM-1 β-lactamase, three active site loops were randomized separately (Palzkill and Botstein, 1992). These separate libraries yielded the key E104K and G238S mutations, resulting in four- and eight-fold increases in cefotaxime resistance, respectively. In combination, however, these two changes result in a 500-fold increase. Although shuffling quickly yielded this double mutation (in addition to other mutations that provided a further sixty-four-fold improvement), the libraries constructed in separate loops based on rational design

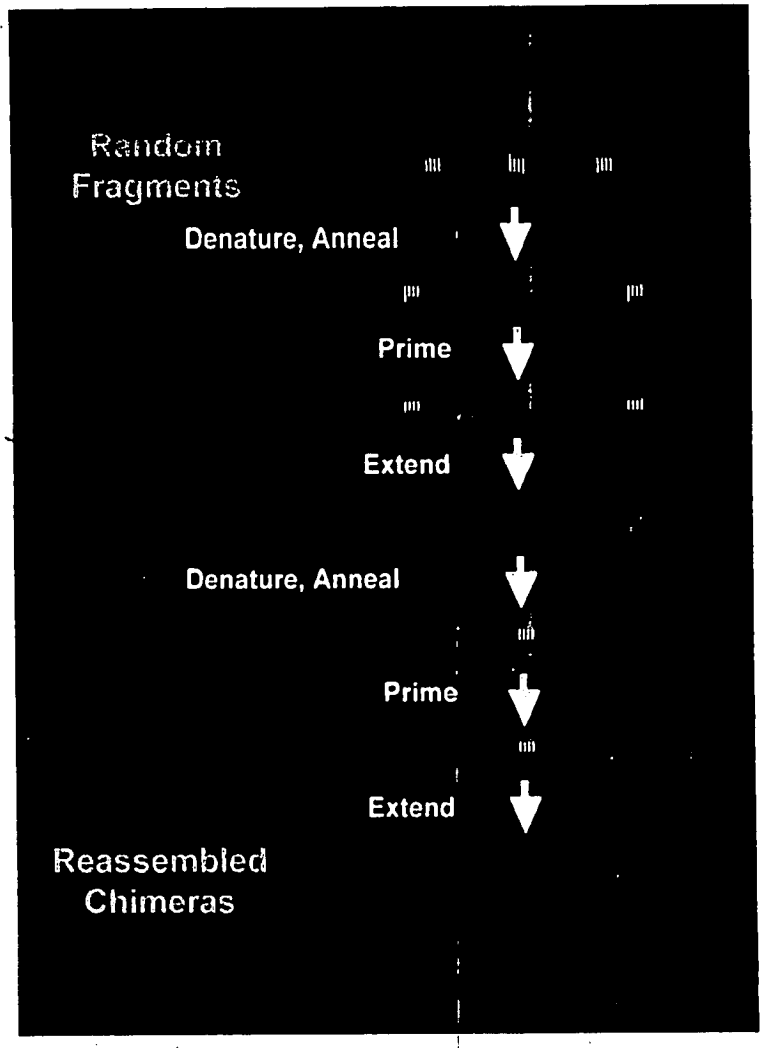

FIG. 1. DNA shuffling by fragmentation and reassembly. A pool of DNA sequences is randomly fragmented (e.g., by treatment with DNase I). The gene fragments are assembled into a library of full-length chimeric genes by repeated cycles of denaturation, annealing, and DNA polymerase extension.

could not yield such a combination mutant. The second advantage of recombination is the purging of the excess of deleterious mutations, which tend to mask the effect of the beneficial mutations. A control experiment using three cycles of error-prone PCR resulted in only a sixteen-fold improvement. Although in principle this method should have yielded the combination of the two key mutations, and the 500-fold increase in MIC, the combination was not seen in practice, probably as a result of the much higher rate of deleterious mutations. Reversal of such deleterious mutations by replacement with wild-type sequence via recombination is efficient, but reversal of deleterious mutations by error-prone PCR is very inefficient. This second advantage is also demonstrated in the example of GFP evolution by DNA shuffling (Crameri *et al.*, 1996). All three of the mutations that resulted in a 45-fold increase in fluorescence were present in a single mutant following the first round of mutagenesis. The effect of two subsequent rounds of shuffling was simply to remove three additional mutations by recombination with wild-type sequences within the pool.

The ability of DNA shuffling to generate a large number of new combinations of beneficial mutations that originated on separate DNA molecules is the key accelerator of directed evolution, since the mutations are often additive or even synergistic in effect (Chen and Arnold, 1991; Matsumura *et al.*, 1986; Shaw *et al.*, 1999; Wells, 1990). By contrast, in nonrecombination approaches the single best parent is typically selected as the sole parent to create the mutant library (either by random or rational mutagenesis) that is to be screened in the next cycle; this effectively discards all except the single best mutation, as well as combinations of mutations that were painstakingly identified in the previous cycle. Iterated cycles of DNA shuffling of a single starting gene using random point mutations as a source of diversity have been successively used to evolve proteins with enhanced activity (Stemmer, 1994b), altered substrate specificity (Zhang *et al.*, 1997), improved protein folding (Crameri *et al.*, 1996), thermostability (Giver *et al.*, 1998), solvent tolerance (Moore *et al.*, 1997), and resistance to chemical modification (Matsumura *et al.*, 1999). Table I summarizes a selection of published examples of DNA shuffling of a single starting sequence.

Although DNase I fragmentation and reassembly is the most widely used format for DNA shuffling, a number of alternative methods have been developed. All rely on the same underlying principle that the most efficient way to explore all possible combinations and permutations of sequences is by recombination. Two proved alternative methods of DNA shuffling are the staggered extension process or StEP (Zhao *et al.*, 1998) and *in vivo* shuffling in *Saccharomyces cerevisiae* (Cherry *et al.*, 1999). Shuffling by random primers (Shao *et al.*, 1998) is not currently widely practiced.

TABLE I
Systems Improved by DNA Shuffling of a Single Starting Sequence

System	Comments	Reference
Single Proteins		
TEM-1 β-lactamase	3 cycles of shuffling and 2 cycles of backcrossing, 32,000-fold increase in antibiotic resistance	Stemmer, 1994
β-galactosidase	7 cycles, 66-fold increase in fucosidase specific activity, 1000-fold increase in substrate specificity	Zhang *et al.*, 1997
Green fluorescence protein	3 cycles, 45-fold improvement in fluorescence as a result of improved protein folding	Crameri *et al.*, 1996
Human antibody	8 cycles of shuffling and 2 cycles of backcrossing, >440-fold increase in avidity	Crameri *et al.*, 1996
Mouse antibody	100-fold increase in expression level	Crameri *et al.*, 1996
Pathways		
Arsenate degradation pathway	3 cycles, 40-fold improvement in arsenate resistance	Crameri *et al.*, 1997

StEP is a PCR-like reaction consisting of a mixture of full-length templates with different beneficial mutations and flanking oligonucleotide primers. In contrast to PCR, StEP employs an extremely abbreviated polymerization step to generate partially extended fragments, which undergo template switching during subsequent cycles of the fragment reassembly reaction. Although StEP can be carried out in a single tube, crossover frequencies are limited by the rapidity with which cycles can be performed.

In vivo DNA shuffling in *S. cerevisiae* uses the cell's highly efficient double-strand DNA break repair pathway to obtain recombination. Yeast cells are co-transformed with a linearized vector and a series of overlapping DNA fragments (e.g., restriction fragments) that together comprise the target sequence. Vector replication requires recircularization of the fragments and vector by a series of *in vivo* recombination events in the homologous overlapping regions. As little as 15 to 30 bp of contiguous identity is all that is required for recombination in yeast (Manivasakam *et al.*, 1995).

Although we have developed and are continuing to develop a range of alternative recombination formats (*in vitro, in vivo,* and combination methods), the *in vitro* methods such as fragmentation and reassembly are currently preferred for most applications due to their versatility and control.

2. *Recombination of Multiple Homologs from Natural Diversity by DNA Shuffling*

The diversity of sequence variations that exists in natural populations of organisms is likely to be very old and highly stable. It is a reservoir of diversity that has proved itself functional and useful. Although this "proved" diversity originated from random mutagenesis, it is much more conservative. For example, introduction of random single amino-acid mutations into a typical protein leads to a high level of nonfunctional "knockout" mutants, yet in dog breeding where one is shuffling two dog genomes that can differ by over a million different single base mutations, most of the puppies that are born are fully functional, suggesting high library quality. Such a result could only be obtained if this large pool of natural diversity was extremely conservative. The fact that such a large fraction of the new combinations of this diversity are so well tolerated suggests that perhaps many of the mutations have already been selected for function and permutability. By contrast, ten generations of randomly mutating dogs will produce a very sick dog rather than a useful new breed.

On the level of the individual gene as well as of the entire organism, changes have been accumulated over millions of years. In the example given above, only genes that function within the context of an entire functional dog have persisted. Similarly, for a single gene, nature has only maintained sequences that are functional within the sequence and structural context of the entire protein and within the complex environment of the whole cell. However, together with the sequence divergence, there has been divergence in a range of properties of gene families. For example, enzymes have evolved to function in diverse physical and chemical conditions (Narinx *et al.*, 1997), to accept new substrates (Scanlan and Reid, 1995), or even to perform fundamentally different chemical reactions (Babbitt *et al.*, 1995). Receptors and ligands have diverged as they co-evolved, maintaining tight binding with each other, but greatly reduced binding to the receptor ligand pairs that have co-evolved in other species. Consequently, as with dog breeding, exchanging blocks of these sequences results in a library containing a large proportion of active members, with a high degree of phenotypic diversity. The permutation of natural sequence diversity encoding amino-acid changes

and neutral mutations, as well as deletions and insertions, produces libraries of progeny that are quite different in sequence and in their combinations of characteristics from any of the parents and therefore represents a broad but sparse sampling of sequence space. Sparse sampling is obtained because the exchange of large sequence blocks by shuffling means that neighboring mutants generally differ at multiple amino-acid positions (Fig. 2). Just as the recombination of dog genomes resulting from sexual reproduction produces functional offspring that differ from either parent, so the molecular breeding of the genes encoding a closely related family of proteins results in a library of functional but different molecules. However, unlike classical breeding, molecular breeding is not limited to two parents and thus bypasses natural species barriers. In addition, molecular breeding is fast with a cycle time of days rather than months or years required for a cycle of classical breeding (Table II).

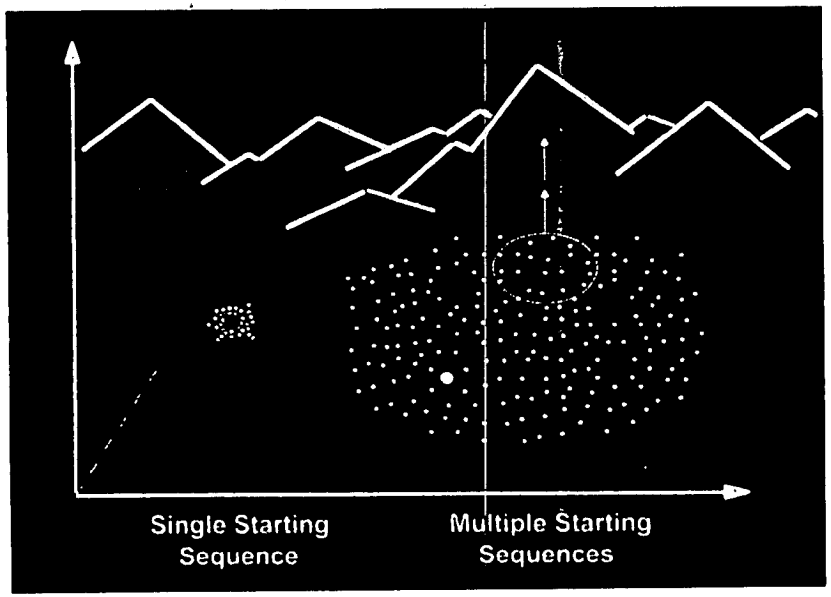

FIG. 2. Searching sequence space by molecular breeding versus random mutagenesis of a single starting sequence. Random mutagenesis yields clones with a few point mutations. The approach is suitable for "hill climbing" to a local performance maximum. Family shuffling yields chimeras that typically have many changes relative to parental sequences and other progeny. At equal library size, the increased sequence diversity results in a sparse sampling of a much greater area of sequence space, allowing much more promising regions to be found and subsequently explored at increased sampling density (Crameri *et al.*, 1998).

TABLE II
Classical Breeding versus Molecular Breeding

Classical breeding	Molecular breeding
Cycle time = years	Cycle time = days
Whole genome	Genes, pathways, genomes
Breed within species	Breed across species
Two parents	One to many parents
Limited control	Multilevel control
Complex selection pressure	Focused selection pressure
Whole plants and animals	Applicable to microbes, cells, whole organisms

The first published example of molecular breeding of natural diversity involved the recombination of four *ampC* β-lactamases that shared 58% to 82% amino-acid identity (Crameri *et al.*, 1998). A library of recombinants was tested for ability to confer increased bacterial resistance to moxalactam, an antibiotic that is poorly degraded by *ampC* β-lactamases. Screening 50,000 members of this library produced a variant that was different from its closest parent at over 25% of its amino acids and conferred 270-fold greater resistance to moxalactam than did the best parent. This compares with an up to eightfold improvement found by single sequence DNA shuffling of any of the four parent genes separately (Fig. 3, see color insert). This enzyme also conferred resistance to a number of other β-lactams.

The shuffling of twenty-six subtilisins and over twenty human interferons are two recent examples that demonstrate the power of molecular breeding to generate high quality and functionally diverse libraries of useful proteins that serve as rich sources of hits when screened for desired properties (Chang *et al.*, 1999; Ness *et al.*, 1999).

a. Interferons. The evolution of pharmaceutical proteins by molecular breeding has recently been demonstrated by Chang *et al.* (Chang *et al.*, 1999). They built a high quality α-interferon library by shuffling more than twenty human α-interferons that shared nucleotide homologies of 85% to 95% (much greater than in the β-lactamase example previously cited). The library was screened as pools of clones for increased antiviral activity, measured by the protection of murine cells against a challenge with lymphocytic choriomeningitis virus, and positive pools were then deconvoluted. Using a total of sixty-eight assays to screen a library of 1672 previously unscreened clones from a single round of recombination, a variant was identified with 135,000-fold greater specific activity than the parent interferon, Hu-IFN-α2a. A second round of shuffling yielded a variant with a 285,000-fold higher specific activity than Hu-IFN-α2a, 185-fold higher specific activity than Hu-IFN-α1, and four-fold more active than the most potent murine interferon, Mu-IFN-α4, despite the fact that the human and murine sequences are only about 65% identical (Fig. 4, see color insert). Importantly, the best three clones were all composed of multiple segments from known human interferons, and contained no new point mutations, suggesting that these clones are much less likely to be immunogenic.

Mice and humans (and their respective cytokines and receptors) have been evolutionarily separated for over 100 million years. Consequently, the murine interferons differ from all of the human interferons at fifty-six to seventy-two amino-acid positions. Nevertheless, it was possible, simply by recombining human sequences, to produce an interferon with an activity greater than that of the natural mouse protein. Sequences that had been "pre-tested" in humans to function as α-interferons contained the information necessary to build a protein that functions well with murine cells. Even though it was not possible to reconstruct the exact sequence of murine α-interferon from the human genes, it was possible to reconstruct its function.

b. Subtilisins. A second example of molecular breeding of a large family of proteins is that of subtilisin. Subtilisins are commercially important serine endoproteases, valued for a range of applications, perhaps most notably as additives to laundry detergents for hydrolysis and solubilization of protein stains (Bott and Betzel, 1996). With annual sales of about $500 million, it is not surprising that subtilisin is one of the best understood proteins and a frequent target for improvement using both structure-based design and random mutagenesis (Ballinger *et al.*, 1996; Bryan *et al.*, 1986; Graycar *et al.*, 1999; Kano *et al.*, 1997; Russel and Fersht, 1987; Wells and Estell, 1988; You and Arnold, 1996). As with most industrial enzymes, incremental improvements in performance are significant. A major challenge in the rational design or directed evolution of industrial enzymes is that performance is not defined by any single property, but by a complex mix of parameters. Although rational design and random mutagenesis can improve single properties, such as thermostability or activity in organic solvent, it is often at the expense of other critical properties (Patkar *et al.*, 1998; Shoichet *et al.*, 1995), making it difficult to obtain an enzyme that is optimized for several of the important performance criteria. Just as multiple traits in plants and animals can be recombined by classical breeding, multiple enzyme properties can be recombined by molecular breeding. Ness *et al.* (Ness *et al.*, 1999) demonstrated this by using DNA shuffling to breed twenty-five subtilisin

Legends for Color Insert

FIG. 3. (a) Comparison of single sequence shuffling and family shuffling of cephalosporinase. (b) Computer model of winning chimera created from the known structure of the *Enterobacter cloacae* protein (Crameri *et al.*, 1998; Lobkovsky *et al.*, 1993). The predicted structure of the α-chain backbone is within an r.m.s deviation of O.766Å from the known structure. The segments derived from *Enterobacter* are shown in blue, those from *Klebsiella* are shown in yellow, and those from *Citrobacter* are shown in green. The thirty-three amino acid point mutations are shown in red. The enzyme differs by 102 amino acids from the *Citrobacter* enzyme, by 142 amino acids from the *Enterobacter* enzyme, by 181 amino acids from the *Klebsiella* enzyme, and by 196 amino acids from the *Yersinia* enzyme.

FIG. 4. Summary of antiviral activities of native and evolved IFN-αs. The antiviral activities of purified protein for native Mu-IFN-αs and Hu-IFN-αs as well as evolved IFN-αs on murine L929 cells are shown. One unit of activity corresponds to half-maximal protection from lethal ECMV viral challenge. Arrows on the right indicate fold improvement of the winning IFN-α (IFN-α-CH2.1) relative to Hu-IFN-α1 and Hu-IFN-α2a (Chang *et al.*, 1999).

FIG. 5. Activities of 654 active clones from the shuffled subtilisin library compared to twenty-six parents. Relative activities of each clone in five screens are plotted as concentric circles. Each color represents one of the five screening conditions: pH5.5 (orange), pH7.5 (blue), pH10 (dark red), thermostability (yellow), and activity in 35% DMF at pH 7.5 (green). The area of the circle is proportional to the activity in the five assays relative to the best parent in each assay.

gene fragments obtained from a panel of mesophilic *Bacillus* isolates with the full-length gene for Savinase, a leading industrial protease (Graycar *et al.*, 1992; Hastrup *et al.*, 1989). The diversity of subtilisins used was much greater than that used in the interferon and β-lactamase examples. Pairwise identities of the DNA sequences were as low as 56.4% (protein sequences homology as low as 63.7%). A small library of 654 active clones was screened for thermostability, solvent stability, and pH dependence (at pH5, pH7.5, and pH10), three properties that are of commercial importance for subtilisin and of general concern for other industrial enzymes and biocatalysts.

The vast array of functional diversity generated in this experiment is shown in Figure 5 (see color insert). The frequency of improved clones ranged from 4% to 12% of the active library in any single parameter. In addition, the diversity of combinations of properties ranged well beyond the properties of the parental enzymes. Sequence analysis of some of the best performing clones under each set of conditions revealed that variants with similar properties could be encoded by very different sequences. Thermostability, for example, could be conferred by any one of at least three different genetic elements. In many applications, a

(a)

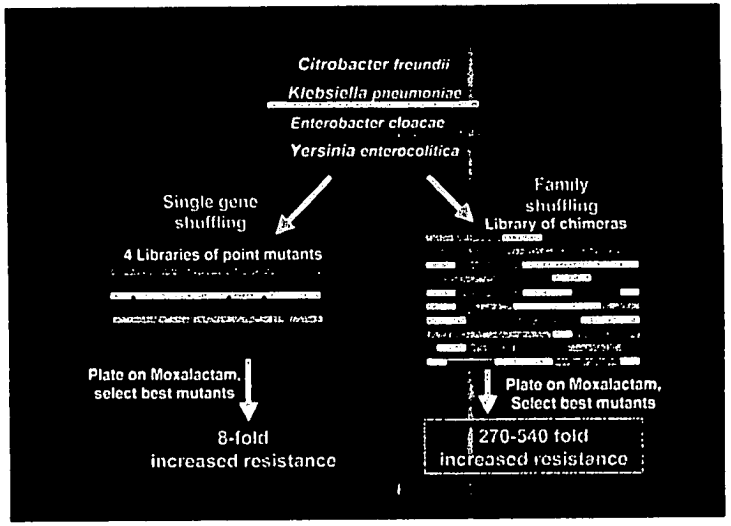

FIG. 3a

(b)

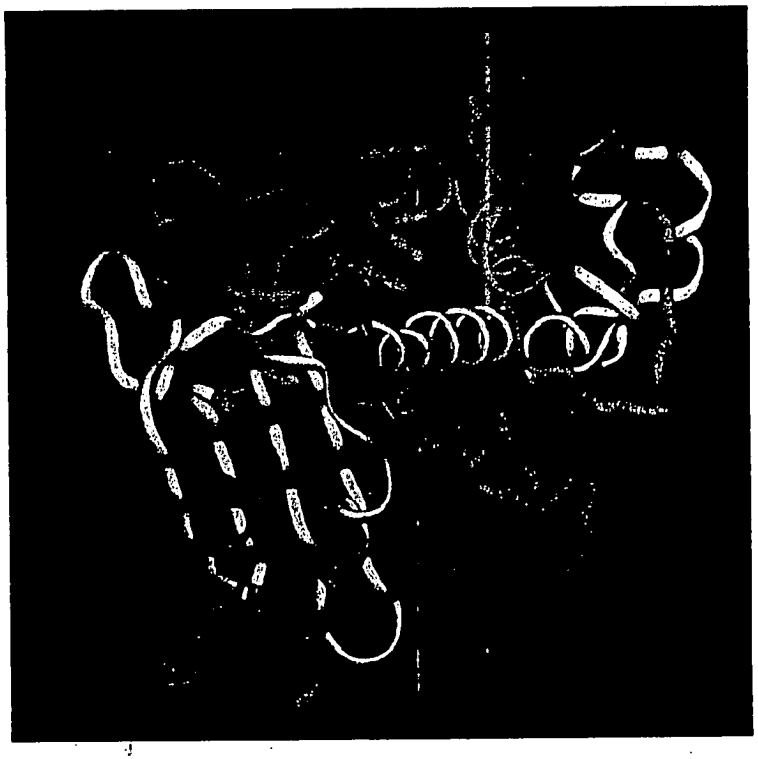

FIG. 3b

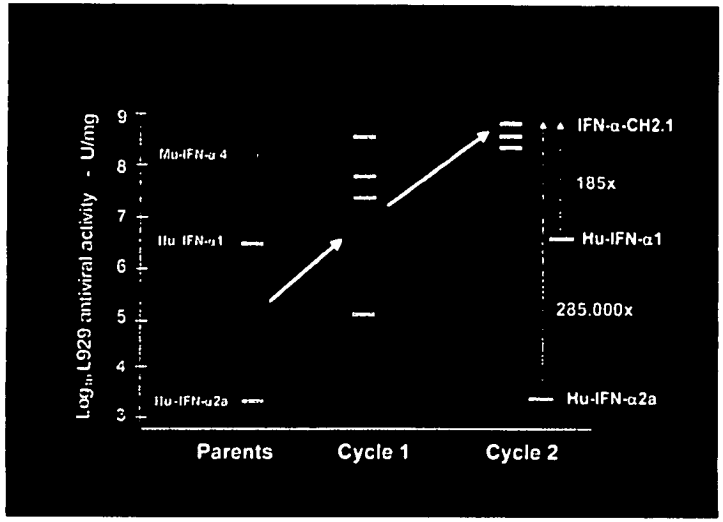

FIG. 4

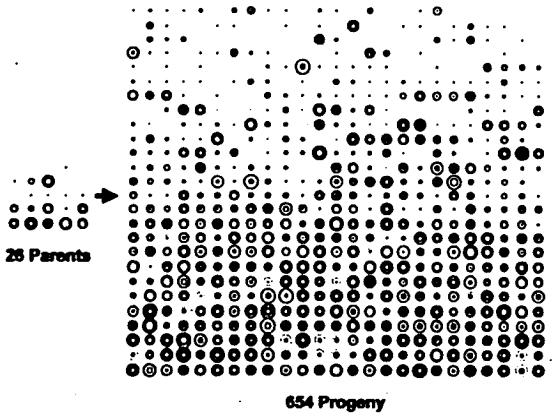

FIG. 5

family of natural sequences is known, but for historical reasons all of the characterization has focused on only one of the family members, typically one whose structure has been determined. Molecular breeding allows less-well characterized sequence homologs (and even partial or inactive sequences) to be incorporated into the breeding pool in any molar ratio desired. The screening of pluripotent enzyme libraries generated by molecular breeding of a handful of homologs provides an economical alternative to rational design or bioprospecting for leads that meet the multiple parameters required for commercialization. In addition, the ability of molecular breeding to demarcate functional sequence elements is likely to be a valuable tool for building structure-function databases for guiding protein design in the future. Table III summarizes a selection of published examples of molecular breeding of multiple related sequences.

IV. Screening Is Key

As Darwin observed, the only difference between the breeding of domestic plants and animals and the evolution of wild organisms is how the selection is applied. In nature, adaptation occurs in response to the environment: Diverse ecological niches give rise to a diversity of organisms to exploit them. Organisms undergo a very low level of random mutagenesis and those mutations that confer a competitive advantage (such as the ability to utilize a new nutrient source, survive at a higher or lower temperature, or kill a neighbor) are maintained in organisms that consequently grow and colonize a new niche.

TABLE III
Systems Improved by Molecular Breeding of Homologous Sequences

System	Comments	Reference
Cephalosporinase	4 sequences, 58%–82% DNA identity, 1 round, 270-540-fold increase in antibiotic resistance	Crameri, Raillard *et al.*, 1998
Thymidine kinase	2 sequences, 78% DNA identity, 4 rounds, 32-16,000-fold decrease in levels of AZT required to sensitize *Escherichia coli*	Christians, Scapozza *et al.*, 1999
α-interferon	>20 sequences, 85%–95% DNA identity, 2 rounds, 185–285,000-fold improvement in specific activity	Chang, Chen *et al.*, 1999
Subtilisin	26 sequences, 56%–99% DNA identity, 1 round, up to 4-fold improvement in 5 properties	Ness, Welch *et al.*, 1999

In classical breeding, the criteria for survival are altered to favor human needs and many of the restraints of natural selection are removed. For example, with people to protect them from predators, cattle are no longer subject to all the rigors of competing in the wild. Humans select and breed from those animals with the highest milk and meat productivity, while alertness and aggression are not only no longer required, but are in fact characteristics that detract from the "performance" of a large domesticated mammal. In a similar way, enzymes used in industrial processes need not be constrained by what is useful to an organism in the wild. For example, a property like product inhibition, which is a critical function of cellular economics, is no longer desirable. The protein economics attempts to subvert the function of the enzyme to be maximally productive and stable under conditions dictated by bioprocess engineers. Like breeders, protein engineers select those enzymes that perform best under the desired conditions.

The challenge of protein design by molecular breeding is the formulation of a screen that precisely emulates the final process conditions. This can be difficult to do in a high throughput format. A powerful approach is the employment of multitiered screens that sample decreasing numbers of clones with increasing scrutiny, ultimately ending with a handful of variants that are tested in the final process. Variants with improved performance under process conditions are carried forward to the next cycle of alteration and screening.

All evolutionary screens require some way to link phenotype with genotype. Recent technologies for linking genotype and phenotype have expanded the accessible library size by many orders of magnitude. For example, cell surface display (Daugherty *et al.*, 1998; Daugherty *et al.*, 1999; Georgiou *et al.*, 1997), phage and virus display (Hodits *et al.*, 1995; Smith *et al.*, 1998; Winter *et al.*, 1994), and ribosome display (Hanes *et al.*, 1999; Hanes *et al.*, 1998; Hanes and Pluckthun, 1997; Mattheakis *et al.*, 1994; Roberts and Szostak, 1997) provide access to libraries of 10^9–10^{14} variants. Unfortunately, screening these libraries remains limited to primarily affinity enrichment. For this reason applications have generally been limited to the identification of polypeptides (mostly antibodies and peptides) that bind tightly to a desired ligand. Although creative exceptions have been reported (Baca *et al.*, 1997; Janda *et al.*, 1994; Smiley and Benkovic, 1994), catalysis or intracellular function have not been conveniently addressed, nor has the technology realized general application such as screening for catalysis.

In some situations it is possible to develop a way to use genetic selection to identify mutants (Black and Loeb, 1993; Naki *et al.*, 1998). By coupling gene function to cell survival (e.g., the acquisition of an essential nutrient, the destruction of a toxic compound, or the ability to activate or complement an existing host metabolic pathway), up to about 10^{12} variants (using combinatorial infection) can be tested. However, the approach is limited by the fact that cells under selective pressure often find unexpected ways to grow (e.g., via genetic reversion or activation of cryptic functions). In addition, selections often have a limited dynamic range. Moreover, it is difficult to distinguish between the specific activity of an enzyme and an increase in its expression level. In general, one must evaluate promising variants from several different angles to avoid undesirable solutions, and must confirm that a phenotype is linked to the gene of interest (i.e., that plasmids from survivors confer viability to a naive cell). Perhaps the main limitation is that a genetic selection for a particular problem is not always obvious. When available, selections are most useful as the first tier ("filter") of a multitiered screening program, which needs to be followed by subsequent, lower throughput but higher veracity screens to evaluate the positive clones.

Fluorescence-based cell sorting also allows screening of large numbers of variants (10^5–10^9). In general, this requires that a fluorescent product is formed and then retained within the cell. An elegant example is the evolution of a P450 enzyme to use hydrogen peroxide in place of the normal NADH cofactor in the hydroxylation of aromatic substrates (Joo *et al.*, 1999). Cells containing active horseradish peroxidase were transformed with a P450 library. Hydroxylated aromatic compounds were linked by the peroxidase to form fluorescent compounds that could then be detected by FACS or digital imaging. The intensity of the fluorescence increased with activity of the P450. In addition, different hydroxylation products resulted in different fluorescence spectra, so that an indication of regioselectivity could be obtained.

Agar plate screens allow the rapid analysis of up to 10^6 different colonies. Screenable phenotypes include enzymes that give rise to a color or fluorescence change in a diffusible substrate (Yang, 1994) or that form a halo around a producing cell because of the degradation of a insoluble substrate, such as the proteolysis of casein in agar plates containing skim milk (Cunningham and Wells, 1987). Plate-based bioassays can be used to detect and quantitate the production of a toxic compound (such as an antibiotic) as a zone of killing of an overlaid tester strain around the producing colony. Coupling of an agar plate-based screen with automated colony picking of positive clones provides a powerful first screen in a multi-tiered screening approach (Ness *et al.*, 1999).

In general, assays that quickly analyze an enormous population of variants tend to compromise the characteristics for which the variants are screened, sacrificing veracity and appropriateness for increased throughput. Consequently, while a high capacity assay may be an efficient way to reduce the number of candidates, it is essential to progress to a more accurate screen before doing additional cycles of recombination. Conditions can typically be manipulated in microtitre plates to give a reasonable approximation of the final process conditions. With a robotic system and a simple homogeneous fluorogenic or chromogenic assay, it is possible to test up to 10^6 variants in microtiter plates. In cases demanding a more complicated screen, such as an assay for stereoselectivity (Janes and Kazlauskas, 1997), about 10^4 clones can be screened in microtiter plates. Finally, manufacturing conditions can be even better simulated in flasks, fermentors, or small reactors, using the protein in its final whole-cell or purified protein form. Low throughput but accurate physical methods such as HPLC, mass spectroscopy, and gas chromatography measure catalytic activity accurately. These high veracity methods are useful as the final assay in a round of directed evolution to ensure that the positive variants really have the required activities and properties.

When optimizing a single property in model systems, such as thermostability or tolerance to an organic solvent, simple high throughput screens have proved adequate (Zhao and Arnold, 1997a). However, to address more complex problems, such as the generation of an improved pharmaceutical protein, a much more elaborate screen may be required. In such cases, molecular breeding of multiple sequences from natural diversity is the best way to generate high-quality libraries that cover a very large area of mostly functional sequence space, so that very few variants need to be tested to obtain the required changes. For example, in the α-interferon example, only sixty-eight assays were used to obtain a significant functional improvement in the first cycle of shuffling. The interferon and subtilisin examples previously described show that the libraries created by molecular breeding can be of unusually high quality. This general approach makes it feasible to perform a small number of assays directly in complex conditions that correlate closely with the final commercial application. We foresee screening small, high-quality libraries of clones directly in whole transgenic plants or animals, especially for whole organism traits for which assays cannot otherwise be obtained, such as yield, drought resistance, or disease resistance.

V. Beyond Proteins

A. *Molecular Breeding of Multigene Phenotypes*

Although most of the previously mentioned methods and examples have focused on improvement of the isolated protein product of a single gene, whole cell biocatalysts make up the majority of industrial biocatalysts. Industrial microorganisms effect the multistep conversion of renewable feedstocks to high value chemical products in a single reactor and comprise a multibillion dollar industry. Fermentation products range from commodity chemicals such as ethanol, organic acids, and amino acids, to high-value small molecule pharmaceuticals, protein pharmaceuticals, and industrial enzymes. Similar to enzymes, whole cell biocatalysts isolated from nature seldom demonstrate the required properties to function under the constraints of a commercial process and thus require specific improvements, such as increased yield of desired products, removal of unwanted co-metabolites, improved utilization of inexpensive carbon and nitrogen sources, and adaptation to fermenter conditions. Success in bringing biocatalytic processes to market and competing in those markets relies on the ability to continuously improve the biocatalyst. The scientific and commercial efforts to understand and manipulate specific functions of whole cells have created the disciplines of metabolic engineering and industrial strain improvement.

Current strategies for strain improvement rely on the empirical and iterative modification of fermenter conditions and genetic manipulation of the whole cell biocatalyst. The genetic manipulation of industrial microorganisms has traditionally taken two paths: the rational approach of metabolic engineering and the empirical approach of classic strain improvement. Although years of intensive research have yielded the genetic tools and information database required to attempt the calculated manipulation of a number of established industrial organisms, metabolic engineering suffers from its reliance on the assumptions of a rational approach. Furthermore, the method and experience gained is typically species-specific and not easily transferred to newly discovered or poorly characterized microorganisms. For these reasons, the most widely practiced strategy is classic strain improvement, which employs random point mutagenesis (chemical or UV) of the producing strain and screening for mutants that have improved properties. Classic strain improvement is robust but suffers from limitations that are typical of iterated point mutagenesis (described above). Molecular breeding, which simulates classical breeding, accelerates whole cell improvement

by removing the limitations of metabolic engineering and classical strain improvement.

The strategies of metabolic engineering generally fall into three classes: (1) enhancing the flux through a desired metabolic pathway by amplifying the expression of genes encoding "rate limiting" enzymes and those resistant to feedback inhibition (see Jetten *et al.*, 1994); (2) the introduction of exogenous genes, which convert a metabolite of the host organism to a desirable chemical at a viable yield (see Cameron and Tong, 1993); and (3) decreasing the diversion of chemical precursors by the disruption of genes encoding competing pathways. All these approaches closely resemble structure-based design of proteins in that they rely on a great deal of information and are often limited by invalid assumptions. The interpretation of biological data is dominated by information considered "known" about the system under investigation. This occurs even though the "known" data set is intrinsically incomplete due to the complexity of biological systems. Recent metabolic engineering studies have demonstrated that cellular physiology is extremely robust, and that well-conceived genetic perturbations often result in little or no change in phenotype. Even severe changes to primary metabolism, such as the deletion of the genes encoding pyruvate kinase, have been shown to have negligible effects on primary metabolic fluxes or growth rate in *E. coli* (Sauer *et al.*, 1999). Complex biological systems from single enzymes to whole cells continue to resist rational manipulation but succumb to empirical approaches, such as mutagenesis and screening, which rely on few assumptions.

Each of the three strategies of metabolic engineering has demonstrated value, yet each is limited by its assumptions and the "cut and paste" nature of genetic engineering. Overexpressing gene(s) believed to represent a rate limiting step or eliminating feedback regulation of pathway enzyme(s) can be a productive means of enhancing flux through desired pathways (class 1 above). However, this approach often results in only a small increase in rate since other genes affecting the pathway become rate limiting. The term "rate-limiting step" is misleading since the rate through a metabolic pathway is generally limited by a collection of enzymes rather than a single enzyme step (Fell, 1998). Metabolic networks are tightly controlled and have evolved to prevent the unnecessary buildup of toxic or useless intermediates. Participating enzymes function at similar rates and under similar conditions to avoid these scenarios, and more than a single enzyme in a given pathway may be under feedback regulation. For example, the biosynthetic enzymes of the aspartate derived amino-acid pathway are under multiple levels of regulation (Eikmanns *et al.*, 1993). Therefore, a small increase in metabolic flux resulting from gene overexpression may be accompanied by a buildup of undesired or detrimental intermediates. Molecular breeding of the genes that encode the pathway enzymes followed by screening the resulting libraries for the desired phenotype provides a direct route to unbiased genetic solutions. This approach allows the improvement of the individual components of the system—for example, improving expression balance within the pathway, eliminating feedback inhibition, improving k_{cat} and K_m for the pathway enzymes, and adaptation to the cellular conditions imparted by the bioprocess. This strategy assumes only that a genetic solution exists within the DNA that is shuffled.

The cloning of heterologous genes to generate new metabolic pathways is one of the most powerful methods for generating new biocatalysts (class 2 above), but poor functioning of the cloned genes often hampers the success of this approach. Genes and gene products have adapted to function in the environment of their native hosts, and these environments are specific to the organisms and their ecological niches. For example, enzymes from thermophilic organisms do not function well in mesophilic hosts. Heterologous genes may be poorly expressed and the encoded polypeptides may not fold properly. Basic genetic elements and the identity of the primary metabolites may be similar between different organisms, but the physical and chemical states of the cells can be significantly different. The concentration of metabolites, pH, temperature, and ionic strength will differ, each influencing the optimal performance of an enzyme; further, the structure of macromolecules with which an enzyme might interact will differ, compromising functional interactions. Thus, a metabolic pathway transplanted from one organism to another may not function optimally. Indeed, the cytoplasmic state of a cell under the conditions of fermentation will be different from that experienced in its natural environment, and even a native pathway may not function optimally. Shuffling heterlogous or native genes and screening them for performance under the desired bioprocess conditions provides a means to identify variants of those genes that have adapted to the new cellular environment and are functioning optimally. The ability of DNA shuffling to alter the substrate preference of enzymes also allows one to access promiscuous activities of enzymes and evolve them to function productively in the context of new metabolic pathways.

The deletion of competing pathways is also a productive route to increasing flux through a desired pathway or at least eliminating potential contaminating products (class 3 above) (Hols *et al.*, 1999). However, the removal of a known pathway may be insufficient to divert flux through the desired pathway, since flux may be limited by either the kinetic parameters of the pathway enzymes or by external factors. Further, the

competing pathway may be essential and its elimination may not be an option. The goal is to divert maximal flux down the desired pathway while maintaining only the necessary flux through any competing pathway. Simultaneous shuffling of both pathways should produce an optimal balance in which flux through the desired pathway is maximized, while maintaining the minimal necessary flux through the competing pathway(s) to allow survival. An intrinsic value of the directed evolution approach is that it allows one to find this balance within a complex system. This often is not possible in a straight metabolic engineering "all-or-nothing" strategy.

DNA shuffling has been demonstrated to improve the heterologous expression of proteins, alter the substrate specificity of enzymes, and improve the function and stability of enzymes under a variety of extreme environmental conditions. Improvement of single genes by DNA shuffling results in the alteration of the expression, structure, and function of the gene product. In contrast, improvement of metabolic pathways by DNA shuffling results in the alteration of the individual genes (as above) as well as complex interactions of the gene products with each other and the cellular environment. In this way, DNA shuffling complements the strategies of metabolic engineering and provides access to the complex genetic solutions required of strain improvement goals. Crameri *et al.* demonstrated the productivity of this approach by the evolution of the *Staphylococcus aureus* arsenate resistance operon to impart increased resistance to arsenate in *E. coli* (Crameri *et al.* 1997). The pathway consisted of three genes encoding an arsenate reductase, an arsenite efflux pump, and a regulatory protein. Previous rational work suggested that any improvements required would be found in the arsenate reductase. After three rounds of shuffling and screening, a variant of the operon imparting a resistance to 0.5 M arsenate was identified (a forty-fold improvement). Analysis of the new operon identified two major surprises: most of the thirteen mutations were clustered in the efflux pump with no mutations found within the coding region of the reductase, and the originally episomal plasmid had integrated into the chromosome and this shuffling-dependent integration was shown to contribute a large part of the improvement. These data emphasize the complex and non-intuitive solutions that arise from a directed evolution approach and demonstrate the utility of molecular breeding to optimize the function of a complete metabolic pathway.

B. Genome Shuffling

Long before molecular geneticists began tinkering with the structure and function of proteins and metabolic pathways, researchers were manipulating the performance of industrial microorganisms by classic strain improvement. These classical approaches remain an important part of all strain improvement programs, primarily because they are robust and reproducibly yield new strains with slightly improved phenotype. Although metabolic engineering requires a great deal of information and molecular tools, classical strain improvement requires only a starting organism, a mutagen, and a good screen for improvement. Cellular phenotypes are complex and are influenced by many more genes than those that are recognized as "necessary and sufficient." Although improvements in phenotype may be accessible by variations within a defined set of genes, other elements distributed throughout the genome may have equal or greater influence. Again, an analogy with protein design is relevant. The improvement of enzyme function is most productive when the entire structural gene is targeted as opposed to only those regions known to encode the active site. Similarly, the most productive mode of improving whole cell biocatalysts is through the evolution of the cell's entire genome. The robust nature of classical strain improvement lies in the fact that it is unbiased and can address complex, distributed phenotypes. The superior performance of classical strain improvement over metabolic engineering is a testament to this fact. The limitations to classical strain improvement are the same as those of sequential point mutagenesis—the process is asexual. Improvements are small and one can accumulate only one beneficial mutation at a time. Genome shuffling incorporates recombination into the strain improvement process and thereby significantly accelerates the process. It provides the means to recombine the genomic information from many strains so that the useful alleles from all of them can be combined into a single superior organism. Useful mutations are combined and deleterious mutations are replaced with wild-type sequence. Instead of accumulating a single beneficial genetic event per cycle of mutagenesis and screening, evolution occurs via large leaps by creating complex combinations of multiple mutations.

VI. Concluding Remarks

Our preferred method for molecular breeding involves recombination of homologous genes obtained from nature, in order to permutate the proven diversity. These libraries are high quality (rich in functional sequences) because the variations have been prescreened for function in nature and phenotypically diverse. In rare cases when adequate natural diversity is not available, such as when homologous sequences are not known or if the target is a small segment of a protein, the sequence

diversity must be generated artificially. Typical methods include random mutagenesis of a single DNA sequence followed by screening for the best mutations, various kinds of synthetic oligonucleotide cassette mutagenesis of a small part of a protein, or mutations that were suggested based on molecular modeling of the protein's structure.

However, regardless of the source of variation, recombination by DNA shuffling is the most effective method for creating higher order combinations of previously selected mutations, whether the targets are single genes, pathways, or whole genomes.

Acknowledgments

We thank Lori Giver, Seran Kim, and Phil Patten for critical input; and Mark Welch for the bubble plot of subtilisins.

References

Arkin, A. P., and Youvan, D. C. (1992). "Optimizing nucleotide mixtures to encode specific subsets of amino acids for semi-random mutagenesis." *Bio/technology*, **10**, 297–300.

Arnold, F. (1998a). "Design by directed evolution." *Acc. Chem. Res.*, **31**, 125–131.

Arnold, F. H. (1998b). "When blind is better: protein design by evolution." *Nature Biotechnology*, **16**, 617–618.

Arnold, G. E., and Ornstein, R. L. (1997). "Molecular dynamics study of time-correlated protein domain motions and molecular flexibility: cytochrome P450BM-3." *Biophys J.*, **73**, 1147–1159.

Babbitt, P. C., Mrachko, G. T., Hasson, M. S., Huisman, G. W., Kolter, R., Ringe, D., Petsko, G. A., Kenyon, G. L., and Gerlt, J. A. (1995). "A functionally diverse enzyme superfamily that abstracts the alpha protons of carboxylic acids." *Science*, **267**(5201), 1159–1161.

Baca, M., Scanlan, T. S., Stephenson, R. C., and Wells, J. A. (1997). "Phage display of a catalytic antibody to optimize affinity for transition-state analog binding." *Proc. Natl. Acad. Sci. USA*, **94**(19), 10063–10068.

Ballinger, M. D., Tom, J., and Wells, J. A. (1996). "Furilisin: a variant of subtilisin BPN' engineered for cleaving tribasic substrates." *Biochemistry*, **35**(42), 13579–13585.

Bhosale, S. H., Rao, M. B., and Deshpande, V. V. (1996). "Molecular and industrial aspects of glucose isomerase." *Microbiol Rev.*, **60**(2), 280–300.

Black, M. E., and Loeb, L. A. (1993). "Identification of important residues within the putative nucleoside binding site of HSV-1 thymidine kinase by random sequence selection: analysis of selected mutants in vitro." *Biochemistry*, **32**, 11618–11626.

Black, M. E., Newcomb, T. G., Wilson, H. M., and Loeb, L. A. (1996). "Creation of drug-specific herpes simplex virus type 1 thymidine kinase mutants for gene therapy." *Proc. Natl. Acad. Sci. USA*, **93**, 3525–3529.

Bornscheuer, U. T., Altenbuchner, J., and Meyer, H. H. (1998). "Directed evolution of an esterase for the stereoselective resolution of a key intermediate in the synthesis of epithilones." *Biotechnology and Bioengineering*, **58**, 554–559.

Bott, R., and Betzel, C. (1996). *Subtilisin Enzymes*, Plenum Press, New York.

Bryan, P. N., Rollence, M. L., Pantoliano, M. W., Wood, J., Finzel, B. C., Gilliland, G. L., Howard, A. J., and Poulos, T. L. (1986). "Proteases of enhanced stability: characterization of a thermostable variant of subtilisin." *Proteins*, **1**, 326–334.

Burbank, L., Whitson, J., John, R., Williams, H. S., and Luther Burbank Society. (1914). *Luther Burbank, his methods and discoveries and their practical application*, Luther Burbank Press, New York; London.

Cadwell, R. C., and Joyce, G. F. (1992). "Randomization of genes by PCR mutagenesis." *PCR Methods Appl.*, **2**(1), 28–33.

Cadwell, R. C., and Joyce, G. F. (1994). "Mutagenic PCR." *PCR Methods Appl.*, **3**(6), S136–S140.

Cameron, D. C., and Tong, I.-T. (1993). "Cellular and metabolic engineering. An overview." *Appl. Biochem. Biotechnol.*, **38**, 105–140.

Chang, C. C., Chen, T. T., Cox, B. W., Dawes, G. N., Stemmer, W. P., Punnonen, J., and Patten, P. A. (1999). "Evolution of a cytokine using DNA family shuffling." *Nat. Biotechnol.*, **17**(8), 793–797.

Chen, G., Dubrawsky, I., Mendez, P., Georgiou, G., and Iverson, B. L. (1999). "*In vitro* scanning saturation mutagenesis of all the specificity determining residues in an antibody binding site." *Protein Eng.*, **12**(4), 349–356.

Chen, K., and Arnold, F. H. (1991). "Enzyme engineering for nonaqueous solvents: random mutagenesis to enhance activity of subtilisin E in polar organic media." *Bio/Technology*, **9**, 1073–1077.

Cherry, J. R., Lamsa, M. H., Schneider, P., Vind, J., Svendsen, A., Jones, A., and Pedersen, A. H. (1999). "Directed evolution of a fungal peroxidase." *1999*, **17**, 379–384.

Christians, F. C., Scapozza, L., Crameri, A., Folkers, G., and Stemmer, W. P. C. (1999). "Directed evolution of thymidine kinase for AZT phosphorylation using DNA family shuffling." *Nature Biotechnol.*, **17**.

Crameri, A., Dawes, G., Rodriguez, E., Silver, S., and Stemmer, W. P. C. (1997). "Molecular evolution of an arsenate detoxification pathway by DNA shuffling." *Nature Biotechnology*, **15**, 436–438.

Crameri, A., Raillard, S.-A., Bermudez, E., and Stemmer, W. P. C. (1998). "DNA shuffling of a family of genes from diverse species accelerates directed evolution." *Nature*, **391**, 288–291.

Crameri, A., Whitehorn, E. A., Tate, E., and Stemmer, W. P. C. (1996). "Improved green fluorescent protein by molecular evolution using DNA shuffling." *Nature Biotechnology*, **14**, 315–319.

Cunningham, B. C., and Wells, J. A. (1987). "Improvement in the alkaline stability of subtilisin using an efficient random mutagenesis and screening procedure." *Protein Eng.*, **1**(4), 319–325.

Dahiyat, B. I., and Mayo, S. L. (1997). "De novo protein design: fully automated sequence selection." *Science*, **278**, 82–87.

Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*, John Murray, London.

Daugherty, P. S., Chen, G., Olsen, M. J., Iverson, B. L., and Georgiou, G. (1998). "Antibody affinity maturation using bacterial surface display." *Protein Eng.*, **11**(9), 825–32.

Daugherty, P. S., Olsen, M. J., Iverson, B. L., and Georgiou, G. (1999). "Development of an optimized expression system for the screening of antibody libraries displayed on the *Escherichia coli* surface." *Protein Eng.*, **12**(7), 613–621.

Delagrave, S., Goldman, E. R., and Youvan, D. C. (1993). "Recursive ensemble mutagenesis." *Protein Eng.*, **6**(3), 327–371.

Delagrave, S., and Youvan, D. C. (1993). "Searching sequence space to engineer proteins: exponential ensemble mutagenesis." *Biotechnology (NY)*, **11**(13), 1548–1552.

Diamond, J. M. (1997). *Guns, germs, and steel: the fates of human societies*, W. W. Norton, New York.

Eikmanns, B. J., Eggeling, L., and Sahm, H. (1993). "Molecular aspects of lysine, threonine, and isoleucine biosynthesis in *Corynebacterium glutamicum*." *Antonie Van Leeuwenhoek*, **64**(2), 145–163.

Fell, D. A. (1998). "Increasing the flux in metabolic pathways: A metabolic control analysis perspective." *Biotechnol Bioeng*, **58**(2-3), 121–124.

Forrest, S. (1993). "Genetic algorithms: principles of natural selection applied to computation." *Science*, **261**, 872–878.

Georgiou, G., Stathopoulos, C., Daugherty, P. S., Nayak, A. R., Iverson, B. L., and Curtiss, R., 3rd. (1997). "Display of heterologous proteins on the surface of microorganisms: from the screening of combinatorial libraries to live recombinant vaccines." *Nat. Biotechnol.*, **15**(1), 29–34.

Gibson, J. M. (1989). "Simulated evolution and artificial selection." *BioSystems*, **23**, 219–229.

Giver, L., Gershenson, A., Freskgard, P. O., and Arnold, F. H. (1998). "Directed evolution of a thermostable esterase." *Proc. Natl. Acad. Sci. USA*, **95**, 12809–12813.

Graycar, T., Knapp, M., Ganshaw, G., Dauberman, J., and Bott, R. (1999). "Engineered *Bacillus lentus* subtilisins having altered flexibility." *J. Mol. Biol.*, **292**(1), 97–109.

Graycar, T. P., Bott, R. R., Caldwell, R. M., Dauberman, J. L., Lad, P. J., Power, S. D., Sagar, I. H., Silva, R. A., Weiss, G. L., Woodhous, L. R., and Estell, D. A. (1992). "Altering the proteolytic activity of stubtilisin through protein engineering." Enzyme Engineering XI, D. S. Clark and D. A. Estell, eds., The New York Academy of Sciences, New York, 71–79.

Hagedorn, S., and Kaphammer, B. (1994). "Microbial biocatalysis in the generation of flavor and fragrance chemicals." *Ann. Rev. Microbiol*, **48**, 773–800.

Hanes, J., Jermutus, L., Schaffitzel, C., and Pluckthun, A. (1999). "Comparison of *Escherichia coli* and rabbit reticulocyte ribosome display systems." *FEBS Lett*, **450**(1–2), 105–110.

Hanes, J., Jermutus, L., Weber-Bornhauser, S., Bosshard, H. R., and Pluckthun, A. (1998). "Ribosome display efficiently selects and evolves high-affinity antibodies *in vitro* from immune libraries." *Proc. Natl. Acad. Sci. USA*, **95**(24), 14130–14135.

Hanes, J., and Pluckthun, A. (1997). "*In vitro* selection and evolution of functional proteins by using ribosome display." *Proc. Natl. Acad. Sci. USA*, **94**(10), 4937–4942.

Hastrup, S., Branner, S., Norris, F., Petersen, S. B., Norskov-Lauridsen, L., Jensen, V. J., and Aaslyng, D. (1989). "Mutated subtilisin genes." PCT Patent Appl. WO 8906279, Novo Industries, Denmark.

Hodits, R. A., Nimpf, J., Pfistermueller, D. M., Hiesberger, T., Schneider, W. J., Vaughan, T. J., Johnson, K. S., Haumer, M., Kuechler, E., Winter, G., *et al.* (1995). "An antibody fragment from a phage display library competes for ligand binding to the low density lipoprotein receptor family and inhibits rhinovirus infection." *J. Biol. Chem.*, **270**(41), 24078–24085.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*, Univ. of Michigan Press, Ann Arbor, MI.

Hols, P., Kleerebezem, M., Schanck, A. N., Ferain, T., Hugenholtz, J., Delcour, J., and de Vos, W. M. (1999). "Conversion of *Lactococcus lactis* from homolactic to homoalanine fermentation through metabolic engineering." *Nat. Biotechnol.*, **17**(6), 588–592.

Janda, K. D., Lo, C. H., Li, T., Barbas, C. F., 3rd, Wirsching, P., and Lerner, R. A. (1994). "Direct selection for a catalytic mechanism from combinatorial antibody libraries." *Proc. Natl. Acad. Sci. USA*, **91**(7), 2532–2536.

Janes, L. E., and Kazlauskas, R. J. (1997). "Quick E. A fast spectrophotometric method to measure the enatntioselectivity of hydrolases." *J. Org. Chem.*, **62**, 4560–4561.

Jespers, L., Jenne, S., Lasters, I., and Collen, d. (1997). "Epitope mapping by negative selection of randomized antigen libraries displayed on filamentous phage." *J. Mol. Biol.*, **269**, 704–718.

Jetten, M. S., Follettie, M. T., and Sinskey, A. J. (1994). "Metabolic engineering of *Corynebacterium glutamicum*." *Ann. NY Acad. Sci.*, **721**, 12–29.

Joo, H., Lin, Z., and Arnold, F. H. (1999). "Laboratory evolution of peroxide-mediated cytochrome P450 hydroxylation [see comments]." *Nature*, **399**(6737), 670–673.

Kano, H., Taguchi, S., and Momose, H. (1997). "Cold adaptation of a mesophilic serine protease, subtilisin, by *in vitro* random mutagenesis." *Appl. Microbiol. Biotechnol.*, **47**, 46–51.

Kelly, K. (1994). *Out of Control: the rise of neo-biological civilization*, Addison-Wesley, Menlo Park, California.

Kobayashi, M., Nagasawa, T., and Yamada, H. (1992). "Enzymatic synthesis of acrylamide: a success story not yet over." *Trends Biotechnol.*, **10**(11), 402–408.

Kunkel, T. A., Bebenek, K., and McClary, J. (1991). "Efficient site-directed mutagenesis using uracil-containing DNA." *Methods Enzymol.*, **204**, 125–139.

Lobkovsky, E., Moews, P. C., Liu, H., Zhao, H., Frere, J. M., and Knox, J. R. (1993). "Evolution of an enzyme activity: crystallographic structure at 2-A resolution of cephalosporinase from the *ampC* gene of *Enterobacter cloacae* P99 and comparison with a class A penicillinase." *Proc. Natl. Acad. Sci. USA*, 90(23), 11257–11261.

Low, N. M., Holliger, P. H., and Winter, G. (1996). "Mimicking somatic hypermutation: affinity maturation of antibodies displayed on bacteriophage using a bacterial mutator strain." *J. Mol. Biol.*, **260**, 359–368.

Manivasakam, P., Weber, S. C., McElver, J., and Schiestl, R. H. (1995). "Micro-homology mediated PCR targeting in *Saccharomyces cerevisiae*." *Nucleic Acids Res.*, **23**(14), 2799–2800.

Matic, I., Rayssiguier, C., and Radman, M. (1995). "Interspecies gene exchange in bacteria: the role of SOS and mismatch repair systems in evolution of species." *Cell*, **80**, 507–515.

Matsumura, I., Wallingford, J. B., Surana, n. K., Vize, P. D., and Ellington, A. D. (1999). "Directed evolution of the surface chemistry of the reporter enzyme β-glucuronidase." *Nature Biotechnology*, **17**, 696–701.

Matsumura, M., Yasumura, S., and Aiba, S. (1986). "Cumulative effect of intragenic amino-acid replacements on the thermostability of a protein." *Nature*, **323**, 356–358.

Mattheakis, L. C., Bhatt, R. R., and Dower, W. J. (1994). "An *in vitro* polysome display system for identifying ligands from very large peptide libraries." *Proc. Natl. Acad. Sci. USA*, **91**(19), 9022–9026.

Minshull, J., and Stemmer, P. (1999). "Protein evolution by molecular breeding." *Curr. Opin. Chem. Biol.*, **3**(3), 284–290.

Moore, J. C., Jin, H. M., Kuchner, O., and Arnold, F. H. (1997). "Strategies for the *in vitro* evolution of protein function: enzyme evolution by random recombination of improved sequences." *J. Mol. Biol.*, **272**(3), 336–347.

Muller, H. J. (1964). "The relation of recombination to mutational advance." *Mutation Research*, **1**, 2–9.

Naki, D., Paech, C., Granshaw, G., and Schellenberger, V. (1998). "Selection of a subtilisin-hyperproducing *Bacillus* in a highly structured environment." *Appl. Microbiol. Biotechnol.*, **49**, 290–294.

Narinx, E., Baise, E., and Gerday, C. (1997). "Subtilisin from psychrophilic antarctic bacteria: characterization and site-directed mutagenesis of residues possibly involved in the adaptation to cold." *Protein Engineering*, **10**, 1271–1279.

Nedwin, G. (1997). "Using enzymes as benign substitutes for synthetic chemicals and harch conditions in industrial processes." Biotechnology in the Sustainable Environment, G. Sayler, J. Sanseverino, and K. Davis, eds., Plenum Press, New York, 13–32.

Ness, J. E., Welch, M., Giver, L., Bueno, M., Cherry, J. R., Borchert, T. V., Stemmer, W. P., and Minshull, J. (1999). "DNA shuffling of subgenomic sequences of subtilisin." *Nat. Biotechnol.*, **17**(9), 893–896.

Ogunseitan, O. A. (1995). "Bacterial genetic exchange in nature." *Science Progress*, **78**(3), 183–204.

Palzkill, T., and Botstein, D. (1992). "Identification of amino acid substitutions that alter the substrate specificity of TEM-1 beta-lactamase." *J. Bacteriol.*, **174**, 5237–5243.

Patkar, S., Vind, J., Kelstrup, E., Christensen, M. W., Svendsen, A., Borch, K., and Kirk, O. (1998). "Effect of mutations in *Candida antarctica* B lipase." *Chem. Phys. Lipids*, **93**, 95–101.

Patten, P. A., Howard, R. J., and Stemmer, W. P. C. (1997). "Applications of DNA shuffling to pharmaceuticals and vaccines." *Curr. Opin. Biotechnol.*, **8**, 724–733.

Reidhaar-Olson, J. F., Bowie, J. U., Breyer, R. M., Hu, J. C., Knight, K. L., Lim, W. A., Mossing, M. C., Parsell, D. A., Shoemaker, K. R., and Sauer, R. T. (1991). "Random mutagenesis of protein sequences using oligonucleotide cassettes." *Methods Enzymol.*, **208**, 564–586.

Roberts, R. W., and Szostak, J. W. (1997). "RNA-peptide fusions for the *in vitro* selection of peptides and proteins." *Proc. Natl. Acad. Sci. USA*. **94**(23), 12297–12302.

Roeder, G. S. (1997). "Meiotic chromosomes: it takes two to tango." *Genes Dev.*, **11**(20), 2600–2621.

Rubingh, D. N. (1997). "Protein engineering from a bioindustrial point of view." *Curr. Opin. Biotechnol.*, **8**, 417–422.

Russel, A. J., and Fersht, A. R. (1987). "Rational modification of enzyme catalysis by engineering surface charge." *Nature*, **328**, 496–500.

Sauer, U., Lasko, D. R., Fiaux, J., Hochuli, M., Glaser, R., Szyperski, T., Wuthrich, K., and Bailey, J. E. (1999). "Metabolic flux ratio analysis of genetic and environmental modulations of *Escherichia coli* central carbon metabolism." *J. Bacteriol.*, **181**(21), 6679–6688.

Scanlan, T. S., and Reid, R. C. (1995). "Evolution in action." *Chem. Biol.*, **2**, 71–75.

Shafikhani, S., Siegel, R. A., Ferrari, E., and Schellenberger, V. (1997). "Generation of large libraries of random mutants in Bacillus subtilis by PCR-based plasmid multimerization." *Biotechniques*, **23**(2), 304–310.

Shao, Z., and Arnold, F. H. (1996). "Engineering new functions and altering existing functions." *Curr. Op. Structural Biol.*, 6, 513–518.

Shao, Z., Zhao, H., Giver, L., and Arnold, F. H. (1998). "Random-priming *in vitro* recombination: an effective tool for directed evolution." *Nucl. Acids Res.*, **26**, 681–683.

Shaw, A., Bott, R., and Day, A. G. (1999). "Protein engineering of α-amylase for low pH performance." *Curr. Opin. Biotechnol.*, **10**(4), 349–352.

Shoichet, B. K., Baase, W. A., Kuroki, R., and Matthews, B. W. (1995). "A relationship between protein stability and protein function." *Proc. Natl. Acad. Sci. USA*, **92**, 452–456.

Smiley, J. A., and Benkovic, S. J. (1994). "Selection of catalytic antibodies for a biosynthetic reaction from a combinatorial cDNA library by complementation of an auxotrophic *Escherichia coli:* antibodies for orotate decarboxylation." *Proc, Natl. Acad. Sci. USA*, **91**(18), 8319–8323.

Smith, G. P., Patel, S. U., Windass, J. D., Thornton, J. M., Winter, G., and Griffiths, A. D. (1998). "Small binding proteins selected from a combinatorial repertoire of knottins displayed on phage." *J. Mol. Biol.*, **277**(2), 317–332.

Smith, J. M., Dowson, C. G., and Spratt, B. G. (1991). "Localized sex in bacteria." *Nature*, **349**, 29–31.

Soong, N.-W., Nomura, L., Pekrun, K., Reed, M., Sheppard, L., Dawes, G., and Stemmer, W. P. C. (2000). "Molecular Breeding of Viruses." *Nature Genetics*. In press.

Spiller, B., Gershenson, A., Arnold, F. H., and Stevens, R. C. (1999). "A structural view of evolutionary divergence." *Proc. Natl. Acad. Sci. USA*, **96**(22), 12305–12310.

Stemmer, W. P. (1994a). "DNA shuffling by random fragmentation and reassembly: *in vitro* recombination for molecular evolution." *Proc. Natl. Acad. Sci. USA*, **91**, 10747–10751.

Stemmer, W. P. (1994b). "Rapid evolution of a protein *in vitro* by DNA shuffling." *Nature*, **370**, 389–391.

Suzuki, M., Christians, F. C., Kim, B., Skandalis, A., Black, M. E., and Loeb, L. A. (1996). "Tolerance of different proteins for amino acid diversity." *Mol. Divers*, **2**(1–2), 111–118.

Taguchi, S., Ozaki, A., and Momose, H. (1998). "Engineering of a cold-adapted protease by sequential random mutagenesis and a screening system." *Appl. Environmental Microbiol.*, **64**(2), 492–495.

van der Meer, J. R., de Vos, W. M., Harayama, S., and Zehnder, A. J. (1992). "Molecular mechanisms of genetic adaptation to xenobiotic compounds." *Microbiol Rev.*, **56**(4), 677–694.

Wells, J. A. (1990). "Additivity of mutational effects in proteins." *Biochemistry*, **29**(37), 8509–8517.

Wells, J. A., and Estell, D. A. (1988). "Subtilisin—an enzyme designed to be engineered." *TIBS*, **13**, 291–297.

Winter, G., Griffiths, A. D., Hawkins, R. E., and Hoogenboom, H. R. (1994). "Making antibodies by phage display technology." *Ann. Rev. Immunol*, **12**, 433–455.

Wong, S. L., Ye, R., and Nathoo, S. (1994). "Engineering and production of streptokinase in a *Bacillus subtilis* expression-secretion system." *Appl. Environ. Microbiol.*, **60**(2), 517–523.

Wu, X. C., Ye, R., Duan, Y., and Wong, S. L. (1998). "Engineering of plasmin-resistant forms of streptokinase and their production in *Bacillus subtilis*: streptokinase with longer functional half-life." *Appl. Environ. Microbiol.*, **64**(3), 824–829.

Yang, M. M. (1994). "Digital imaging spectroscopy of microbial colonies." *Am. Biotechnol Lab*, **12**(6), 18–20.

You, L., and Arnold, F. H. (1996). "Directed evolution of subtilisin E in *Bacillus subtilis* to enhance total activity in aqueous dimethylformamide." *Protein Eng.*, **9**, 77–83.

Zhang, J.-H., Dawes, G., and Stemmer, W. P. C. (1997). "Directed evolution of a fucosidase from a galactosidase by DNA shuffling and screening." *Proc. Natl. Acad. Sci. USA*, **94**, 4505–4509.

Zhao, H., and Arnold, F. (1997a). "Combinatorial protein design: strategies for screening protein libraries." *Current Opinion in Structural Biology*, **7**, 480–485.

Zhao, H., and Arnold, F. H. (1997b). "Functional and nonfunctional mutations distinguished by random recombination of homologous genes." *Proc. Natl. Acad. Sci. USA*, **94**, 7997–8000.

Zhao, H., and Arnold, F. H. (1997c). "Optimization of DNA shuffling for high fidelity recombination." *Nucleic Acids Res.*, 25(6), 1307–1308.

Zhao, H., Giver, L., Shao, Z., Affholter, J. A., and Arnold, F. H. (1998). "Molecular evolution by staggered extension process (StEP) *in vitro* recombination." *Nature Biotechnol*, **16**, 258–261.

ANALYSIS OF LARGE LIBRARIES OF PROTEIN MUTANTS USING FLOW CYTOMETRY

By GEORGE GEORGIOU

Department of Chemical Engineering and Institute for Cell and Molecular Biology, University of Texas, Austin, Texas 78712

I. Introduction

One of the most intriguing problems in directed protein evolution is determining the optimal strategy for exploring the sequence space and isolating gain-of-function, change-of-function, or stability mutants. The most widely adopted experimental strategy is the iterative search of libraries containing low rates of random nucleotide substitutions (Arnold, 1998). This approach has been dictated by both experimental limitations as well as theoretical considerations: Assaying for enzyme function is generally tedious and represents the rate limiting step in library screening. Until recently, the number of independent clones that could be assayed for enzymatic function by liquid or plate assays was limited to around 5×10^5 clones (Joo *et al.*, 1999). If a low rate of mutagenesis is used, a large fraction of all the possible amino-acid substitutions ("sequence space") may be represented in a library that is still small enough to be screened by conventional agar plate or microtiter well assays. In addition, a low rate of mutagenesis is considered necessary to maintain the fraction of deleterious mutations at a tolerable level (Kuchner and Arnold, 1998).

The iterative screening of relatively small libraries of mutants with a low frequency of nucleotide substitutions has proved to be extremely effective for the functional improvement of numerous proteins and has literally changed the way we think about protein design (Arnold, 2000). However, as high throughput screening methodologies are becoming

ADVANCES IN PROTEIN CHEMISTRY

EDITED BY

FREDERIC M. RICHARDS
Department of Molecular Biophysics
and Biochemistry
Yale University
New Haven, Connecticut

DAVID S. EISENBERG
Department of Chemistry and Biochemistry
University of California, Los Angeles
Los Angeles, California

PETER S. KIM
Department of Biology
Massachusetts Institute of Technology
Whitehead Institute for Biomedical Research
Howard Hughes Medical Institute Research Laboratories
Cambridge, Massachusetts

VOLUME 55

Evolutionary Protein Design

EDITED BY

FRANCES H. ARNOLD
California Institute of Technology
Pasadena, California

ACADEMIC PRESS
San Diego London Boston New York
Sydney Tokyo Toronto